# Reliability and stability of the Roland Morris Disability Questionnaire

Brouwer, S; Kuijer, W; Dijkstra, PU; Goeken, LNH; Groothoff, JW; Geertzen, JHB

*Published in:*
Disability and Rehabilitation

*DOI:*
10.1080/09638280310001639713



*Publication date:*
2004

DISABILITY AND REHABILITATION, 2004; VOL. 26, NO. 3, 162–165

Taylor & Francis
healthsciences

# Reliability and stability of the Roland Morris Disability Questionnaire: intra class correlation and limits of agreement

S. BROUWER†,*, W. KUIJER†, P.U. DIJKSTRA†‡, L.N.H. GÖEKEN†¶, J.W. GROOTHOFF§ and J.H.B. GEERTZEN†§

† Centre for Rehabilitation, University Hospital Groningen, The Netherlands
‡ Department of Oral and Maxillofacial Surgery, University Hospital Groningen, The Netherlands
§ Northern Centre for Health Care Research, University of Groningen, The Netherlands
¶ Institute for Movement Sciences, University of Groningen, The Netherlands



*Accepted for publication*: October 2003

**Abstract**

*Purpose*: To analyse test-retest reliability and stability of the Dutch language version of the Roland Morris Disability Questionnaire (RMDQ) in a sample of patients ($n = 30$) suffering from Chronic Low Back Pain (CLBP).
*Method*: Patients filled out the Dutch language version of the RMDQ questionnaire twice, before starting the rehabilitation programme, with a 2-week interval. Intra Class Correlations (ICC), (one way random) was used as a measure for reliability and the limits of agreement were calculated for quantifying the stability of the RMDQ. An ICC of 0.75 or more was considered as an acceptable reliability. No criteria for limits of agreement were available. However, smaller limits of agreement indicate more stability because it indicates that the natural variation is small.
*Results*: The Dutch RMDQ showed good reliability, with an ICC of 0.91. Calculating limits of agreement to quantify the stability, a large amount of natural variation ($\pm$ 5.4) was found relative to the total scoring range of 0 to 24.
*Conclusion*: The Dutch RMDQ proves to be a reliable instrument to measure functional status in CLBP patients. However, the natural variation should be taken into account when using it clinically.

## Introduction

Functional status is an important evaluative outcome measure in low back pain rehabilitation.[1, 2] To assess changes in functional status after treatment in patients with low back pain, the Roland Morris Disability Questionnaire (RMDQ) is frequently used.[2–4] The RMDQ is derived from the Sickness Impact Profile, a general health questionnaire.[5]

For an outcome measurement, it is important that the reliability is good and that repeated measures in individuals remain stable over time,[6] in the absence of treatment. In reliability studies of the RMDQ, Pearson correlation coefficient is often used as a measure for reliability.[2, 7, 8] Pearson correlation reflects the extent to which two repeated measures can be fitted by a straight line. The disadvantage of this statistic measure is that repeated measures may differ systematically (statistically), yet correlate highly or perfectly. By contrast, the intra-class correlation coefficient (ICC) assesses not only the strength of correlation, but also if all measures on each subject are identical, and do not differ systematically. Therefore, ICC is preferable over the Pearson correlation to use as measure for reliability. But usually the Pearson coefficient will be higher than the ICC and may be used more often for that reason.

Stability over time, in the absence of treatment, may be influenced by within-patient variance and random errors. These sources of variance may lead to instability or fluctuations on the RMDQ-scale: 'natural variations'.[6] If a person fills out the same questionnaire on two occasions, it is relevant to know what variation in test scores can be expected in the absence of treatment.

* Author for correspondence; Centre for Rehabilitation, University Hospital Groningen, PO Box 30.001, 9700 RB, Groningen, The Netherlands. e-mail: s.brouwer@rev.azg.nl

*Disability and Rehabilitation* ISSN 0963–8288 print/ISSN 1464–5165 online © 2004 Taylor & Francis Ltd
http://www.tandf.co.uk/journals
DOI: 10.1080/09638280310001639713

To investigate this natural variation on the RMDQ-scale, limits of agreement can be calculated according to the method of Bland and Altman.[9] In an individual patient the change due to treatment should exceed these limits of agreement before one can state that the treatment has been effective. Therefore, limits of agreement should be taken into account when using the RMDQ clinically.

The English version of the RMDQ shows good reliability.[1, 7, 10] However, limits of agreement have not been investigated. A validated Dutch language version of the RMDQ is available,[11] but test-retest reliability and limits of agreement have not been investigated previously.

The aim of this study is to investigate the test-retest reliability of the Dutch RMDQ for patients with chronic low back pain (CLBP), using ICC as measure for reliability, as well as to quantify the stability of the RMDQ by calculating limits of agreement.



## Methods

### GENERAL PROCEDURE

Patients with CLBP were recruited from the population who were admitted for rehabilitation treatment of the Centre for Rehabilitation at the University Hospital Groningen. Patients were included in the study if they were between 18 – 65 years of age, still at work, or were less than 1 year out off work due to CLBP. Exclusion criteria were specific low back pain, entirely off work for a year or more, cardiovascular or pulmonary diseases, pregnancy, addiction, and psychopathology. Patients filled out the Dutch language version of the RMDQ, before starting the rehabilitation programme, with a 2-week interval. Time, day and place of assessment were held constant for the two test-sessions. The present study was approved by the Medical Ethical Committee of the University Hospital Groningen.

### POPULATION

Thirty patients (24 male and 6 female) with CLBP participated in this study. All patients were referred for treatment in a rehabilitation centre between May 2000 and April 2001 and agreed to participate. Demographics and medical history were obtained of all patients. The mean age of the patients was 40 years (SD 8.1 year). The duration of low back pain ranged between 5 and 10 years. Patients were off work for a mean of 17 weeks (SD 19.2). Fifteen patients (50%) were receiving financial compensation.

### DUTCH LANGUAGE VERSION OF THE RMDQ

The Dutch language version of the RMDQ is a translation of the original RMDQ.[8] It assesses perceived restrictions in 24 activities of daily living dichotomously. The sum score is calculated by summing the 'yes' answers. The scale ranges from 0 (no disability) to 24 (severe disability).

### DATA ANALYSES

Descriptive statistics were calculated for the total scores of the two test-sessions. Test-retest reliability was determined by means of a paired *t*-test, intra class correlation coefficient (ICC, one way random) for the sum scores. Limits of agreement were used to determine the natural variation for quantifying stability over time.[9, 12] To calculate limits of agreement, a plot of the difference between the two sessions for each patient against the mean of each patient of the two sessions was made. Then the average difference in the two sessions, and the standard deviation of the difference between the two scores ($SD_{change}$) were calculated. Finally, the limits of agreement were calculated, equal to twice the standard deviation. An ICC above 0.75 was considered as good reliability.[13–15] No criteria for interpretation of the limits of agreement were available. However, smaller limits of agreement indicate more stability because it indicates that the natural variation is small. Data analyses were performed using the Statistical Package for Social Sciences (SPSS 10.0).

## Results

Mean of the sum score in the first and second session was respectively 13.0 (SD 4.8) and 12.1 (SD 5.0). The mean difference was 0.83 (SD 2.7) (95% CI of the difference: $-0.2$ to 1.8). The ICC was 0.91 (95% CI: 0.82 to 0.96). Limits of agreement were $\pm$ 5.4 (figure 1).

## Discussion

No systematic differences were found in the sum score of the first and the second session. The reliability of the Dutch RMDQ was good (ICC (one way random) above the criterion of 0.75). Similar results of ICCs of 0.75 or higher were found in many other RMDQ studies.[10, 16–21] However, also considerable lower ICCs were found ranging from 0.42 to 0.66.[21, 22] Most studies with ICC values of $\geqslant 0.75$ used an interval of 1 – 14 days between the two sessions, whereas for the studies with ICCs below 0.75, the interval was 6 weeks or more. Almost

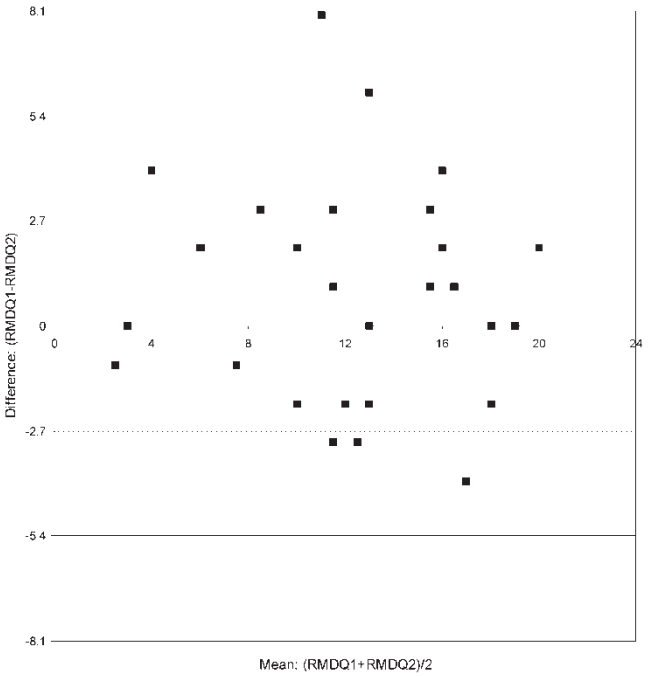


**Figure 1** Difference between RMDQ1 and RMDQ 2 plotted against average of RMDQ1 and RMDQ2.

all studies with a time interval of more than 2 weeks have lower ICCs than the studies with an interval of 2 weeks or less. An explanation of this phenomenon might be that a shorter interval between the two sessions may result in patients remembering the score of the previous session. A larger interval between the two sessions may result in loss of remembering the score of the previous session and change of clinical status in that period. Reliability of functional status questionnaires may be best measured using an interval of 1–2 weeks, a period in which the clinical status is reasonably stable in chronic pain patients.[4] In our study we used an interval of 2 weeks.

Comparing studies using Pearson correlation[2, 7, 8] with studies using ICC[10, 16–21] as a measure of reliability, it appears that the magnitude of Pearson and ICC are similar, i.e., the reliability is good. This suggests that the predominant source of error is due to random variation instead of a systematic difference. Under these circumstances, the Pearson and ICC are very similar.[14]

To quantify stability, we investigated the natural variation by calculating limits of agreement according to the method of Bland and Altman.[9] Despite the good reliability (ICC), the limits of agreement ($\pm$ 5.4) were large relative to the total scoring range of 0 to 24. This means that within person variance or random errors have led to instability in measurement results, approximately 95% of all differences within persons will lie between $\pm$ 5.4. This large amount of natural variation should be taken into account when using the RMDQ clinically. Effects of therapy should exceed the limits of agreement before one can state that the treatment has been effective. *Post-hoc* analysis showed that for all items $\geqslant$ 70% of the scores were the same for the two sessions. Thus, the large amount of natural variation could not be contributed to some specific items.

De Vet *et al*.[6] used the Smallest Real Differences for individuals ($SRD_{individual}$) as measure for quantifying the stability of the RMDQ. Despite the use of different terms, the calculation of the limits of agreement and $SRD_{individual}$ are the same. We found limits of agreement of 5.4 on a scale of 0–24 on the RMDQ, De Vet *et al*.[6] found a $SRD_{individual}$ value of 5.9. Limits of agreement, in our study, were calculated on the basis of scores collected before patients started the intervention, to minimize the possibility that a clinically important change of the construct would occur in the period of data collecting. The study of De Vet *et al*.[6] however, is an intervention study and the $SRD_{individual}$ was calculated on the basis of the scores of a group of patients who rated themselves as not clinically important changed despite the intervention. An estimation of not clinically important changed was obtained by global perceived effect assessed by the patient on a 7-points transition scale (1 = completely recovered, 7 = vastly worsened).[23] The validity of a transition scale however is problematic; it cannot be regarded as a 'golden standard'. Bias may affect subjective assessment of change, patients do not remember correctly how they felt at the beginning of treatment. Moreover, they usually underestimate their initial state, resulting in exaggerating the effect of the program.[14] Classifying patients as 'changed' or 'not-changed' on the basis of these results may be biased.

Limits of agreement can also be used as a cut-off score for change in an intervention study or in daily practice. The cut-off change score determines the minimum change that is considered to be clinically relevant.[24, 25] Based on our results, patients have to change at least 6 points on a scale of 0–24 of the RMDQ to exceed the natural variation and to be judged as having really changed.

Several intervention studies have determined the cut-off score for change of the RMDQ,[4, 21, 26, 27] ranging from 2 to 5 points. These studies underestimate the height of the cut-off score; changes on the RMDQ scale ranging from 2 to 5 points cannot be detected as a clinically relevant change, given the natural variation we found.

## Conclusion

The Dutch RMDQ proves to be a reliable instrument to measure functional status in CLBP patients. However, a large amount of natural variation (±5.4) was found relative to the total scoring range of 0 to 24.

**Acknowledgement**

The authors like to thank Rita Schiphorst-Preuper, Cor Muskee, Willem Jorritsma and Janine Stubbe for their valuable assistance in selecting patients and collecting data. This study was supported by ZonMw grant number 96-06-006.

**References**

1 Deyo R. Comparative validity of the sickness impact profile and shorter scales for functional assessment in low-back pain. *Spine* 1986; **11**: 951–954.
2 Bombardier C. Outcome assessments in the evaluation of treatment of spinal disorders: summary and general recommendations. *Spine* 2000; **25**: 3100–3103.
3 Roland M, Fairbank J. The Roland-Morris Disability Questionnaire and the Oswestry Disability Questionnaire. *Spine* 2000; **25**: 3115–3124.
4 Beurskens AJ, De Vet HC, Koke AJ, van der Heijden GJ, Knipschild PG. Measuring the functional status of patients with low back pain. *Spine* 1995; **20**: 1017–1028.
5 Bergner M, Bobbitt RA, Carter WB, Gilson BS. Sickness impact profile: development and final revision of health status measure. *Medical Care* 1981; **19**: 787–805.
6 De Vet HCW, Bouter LM, Bezemer PD. Reproducibility and responsiveness of evaluative outcome measures. *International Journal of Technology Assessment in Health Care* 2001; **17**(4): 479–487.
7 Jensen MP, Strom SE, Turner J, Romano JM. Validity of the sickness impact profile Roland scale as a measure of dysfunction in chronic pain patients. *Pain* 1992; **50**: 157–162.
8 Roland M, Morris R. A study of the natural history of back pain, part I: development of a reliable and sensitive measure of disability in low back pain. *Spine* 1983; **8**: 141–144.
9 Bland JM, Altman DG. Statistical methods for assessing agreement between two methods of clinical measurement. *Lancet* 1986; **i**: 307–310.
10 Stratford PW, Binkley JM, Riddle DL. Development and initial validation of the back pain functional scale. *Spine* 2000; **25**: 2095–2102.
11 Gommans IHB, Koes BW, van Tulder MW. Validiteit en responsiviteit Nederlandstalige Roland Disability Questionnaire. Vragenlijst naar functionele status bij patiënten met lage rugpijn. *Nederlands Tijdschrift voor Fysiotherapie* 1997; **107**: 28–33.
12 Altman DG, Bland JM. Measurement in medicine: The analysis of method comparison studies. *Statistician* 1983; **32**: 307–317.
13 Lee J, Koh D, Ong CN. Statistical evaluation of agreement between two methods for measuring quantitative variables. *Computers in Biology and Medicine* 1989; **19**: 61–70.
14 Streiner DL, Norman GR. *Health Measurement Scales: A practical Guide to Their Development and Use*. Oxford: Oxford University Press, 1995; 104–127.
15 Tammemagi MC, Frank JW, LeBlanc M, Artsob H, Streiner DL. Methodological issues in assessing reproducibility – A comparative study of various indices of reproducibility applied to repeat elisa serologic tests for lyme disease. *Journal of Clinical Epidemiology* 1995; **48**: 1123–1132.
16 Johansson E, Lindberg P. Subacute and chronic low back pain: reliability and validity of a Swedish version of the Roland and Morris disability Questionnaire. *Scandinavian Journal of Rehabilitation Medicine* 1998; **30**: 139–143.
17 Kopec JA, Esdaile JM, Abrahamowicz M, Abenhaim L, Wood-Dauphinee S, Lamping DL, Williams JI. The Quebec Back pain disability scale: measurement properties. *Spine* 1995; **20**: 341–352.
18 Nusbaum L, Natour J, Ferraz MB, Goldenberg J. Translation, adaptation and validitation of the Roland-Morris questionnaire: Brazil Roland-Morris. *Brazilian Journal of Medical and Biological Research* 2001; **34**: 203–210.
19 Underwood MR, Barnett AG, Vickers MR. Evaluation of two time-specific back pain outcome measure. *Spine* 1999; **24**: 1104–1112.
20 Jacob T, Baras M, Zeev A, Epstein L. Low back pain: reliability of a set of pain measurement tools. *Archives of Physical Medical and Rehabilitation* 2001; **82**: 735–742.
21 Patrick DL, Deyo RA, Atlas SJ, Singer DE, Chapin A, Keller RB. Assessing health-related quality of life in patients with sciatica. *Spine* 1995; **20**: 1899–1908.
22 Davidson M, Keating JL. A comparison of five low back disability questionnaires: reliability and responsiveness. *Physical Therapy* 2002; **82**: 8–24.
23 Beurskens AJ, De Vet HC, Koke AJ. Responsiveness of functional status in low back pain. A comparison of different instruments. *Pain* 1996; **65**: 71–76.
24 Goldsmith CH, Boers M, Bombardier C, Tugwell P. Criteria for clinically important changes in outcomes: Development, scoring and evaluation of rheumatoid arthritis patient and trial profiles. *Journal of Rheumatology* 1993; **20**: 561–565.
25 Wells GA, Tugwell P, Kraag GR, Baker PR, Groh J, Redelmeier DA. Minimum important difference between patients with rheumatoid arthritis: The patient's perspective. *Journal of Rheumatology* 1993; **20**: 557–560.
26 Deyo RA, Centor RM. Assessing the responsiveness of functional scales to clinical change: an analogy to diagnostic test performance. *Journal of Chronical Diseases* 1986; **11**: 897–906.
27 Stratford PW, Binkley JM, Solomon P, Gill C, Finch E. Assessing change over time in patients with low back pain. *Physical Therapy* 1994; **74**: 528–533.